# A Benchmark Study on Sentiment Analysis for Software Engineering Research

MSR 2018

Vivek
North Carolina State University
vvivek@ncsu.edu

## ABSTRACT

Sentiment analysis study on collaborative software forums and tools has become a research trend. For this purpose existing sentiment analysis tools developed for general purpose has been of limited success. It is now understood that sentiments attached to software projects needs to be analysed differently than the general communications. Novielli et al. [2] reports a benchmark study to assess the performance and reliability of three sentiment analysis tools specifically customized for software engineering. They have also stated manual error analysis of the data and classifiers to indicate the quality of the classification.

## KEYWORDS

Sentiment analysis; Communication Channels; Social Software Engineering; NLP

## 1 INTRODUCTION

Sentiment analysis is the study of subjectivity and polarity of emotions as indicative of affected states from textual content. In this study, sentiment analysis tools deal specifically with textual content that relate to software projects. They have replicated the study by Jongeling et al. [1] to assess the performance and reliability of three sentiment analysis tools. Jongeling et al. had established that off-the-shelf sentiment analysis tools (such as NLTK) have been trained on non-technical data, hence they produce unreliable results in the technical domain.

The research questions that the authors have considered are as follows:

- **RQ1: To what extent do different SE-specific sentiment analysis tools agree with emotions of software developers?**
- **RQ2: To what extent do results from different SE-specific sentiment analysis tools agree with each other**

Following three sentiment analysis tools have been selected for analysis: **Senti4SD**, **SentiStrengthSE** and **SentiCR** The performance of the above tools has been compared with the baseline which is represented by SentiStrength. The performance is assessed on Jira and Stack Overflow datasets. For understanding affect of labeling these two approaches can be adopted: model-driven and ad-hoc. As defined by the authors, model-driven annotation is inspired by theoretical models of affect, which are translated into detailed guidelines and are used as a reference for the human raters, after a preliminary training. Jira and Stack Overflow datasets belong to this approach. Whereas, in ad-hoc annotation, the raters are required to provide polarity labels according to their subjective perception of the semantic orientation of the text. Since, the quality of gold standards could not ascertained to one particular approach, so the reliability of sentiment analysis depends on the approach taken. Hence,

- **RQ3: To what extent do the labeling approach (model-driven vs. ad-hoc annotation) has an impact on the performance of SE specific sentiment analysis tools?**

## 2 SENTIMENT ANALYSIS TOOLS

In SentiStrength, each negative word gets a score from -2 to -5 and positive ones from +2 to +5. All neutral ones are -1 and +1. It also accounts for exclamation marks, two emotion words consecutively and emoticons among other rules. Senti4SD is a supervised polarity classifier. It is publicly available with a gold standard of about 4K questions, answers and comments from StackOverflow, manually annotated. Senti4SD leverages a suite of features based on n-grams, sentiment lexicons and semantic features based on word embedding, whose contribution is assessed by the authors through an empirical evaluation leveraging different feature settings. [2]. SentiStrengthSE is an unsupervised classifier that leverages a manually adjusted version of the SentiStrength lexicon and implements ad-hoc heuristics to correct the misclassifications. SentiCR is a supervised sentiment analysis toolkit, specifically trained and evaluated for code review comments. SentiCR has been evaluated using eight supervised algorithms in a 10-fold cross validation setting.

## 3 EVALUATION METRICS

The results have been reported in terms of precision, recall, and f-measure for all the three polarity classes. Micro- and macro- averaged values of precision, recall and f-measures have also been reported. Weighted kappa (Îž) a measure of inter-rater agreement is used to assess agreement with gold labels (RQ1) and the agreement among the three tools (RQ2). This is used to differentiate between ratings provided as positive/negative with neutral ones and positive with negative ones. The former is called as mild disagreement and the latter as strong disagreement. The classifiers are trained on training set that replicates the experimental setting provided in original studies. The test set is used to build the SentiStrength baseline.

*FSS, Fall 2018, Computer Science*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$9999.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

**Table 1: Performance of sentiment analysis tools for model-driven annotations**

| Class | SentiStrength Baseline | | | Senti4SD | | | SentiStrengthSE | | | SentiCR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| (SO)Positive | .89 | **.92** | .90 | **.92** | **.92** | **.92** | .89 | .83 | .86 | .88 | .90 | .89 |
| (SO)Negative | .67 | **.96** | .79 | **.80** | .89 | **.84** | .75 | .79 | .77 | .79 | .73 | .76 |
| (SO)Neutral | **.95** | .64 | .76 | .87 | .80 | **.83** | .75 | .77 | .76 | .79 | **.82** | .80 |
| (SO)Micro-avg. | .82 | .82 | .82 | **.87** | **.87** | **.87** | .80 | .80 | .80 | .82 | .82 | .82 |
| (SO)Macro-avg. | .84 | .84 | .84 | **.86** | **.87** | **.86** | .80 | .80 | .80 | .82 | .81 | .82 |
| (Jira)Positive | .50 | **.91** | .65 | **.76** | .79 | .78 | .69 | .94 | .80 | **.76** | .89 | **.82** |
| (Jira)Negative | .41 | .64 | .50 | .72 | .57 | .64 | .67 | .71 | .69 | **.81** | **.61** | **.70** |
| (Jira)Neutral | .89 | .59 | .71 | .86 | **.89** | .88 | **.92** | .82 | .87 | .89 | **.89** | **.89** |
| (Jira)Micro-avg. | .66 | .66 | .66 | .83 | .83 | .83 | .83 | .83 | .83 | **.85** | **.85** | **.85** |
| (Jira)Macro-avg. | .60 | .71 | .62 | .78 | .75 | .76 | .76 | **.82** | .78 | **.82** | .80 | **.80** |

**Table 2: Agreement of SE-specific tools with manual labelling (a) and with each other (b) for model-driven annotations.**

| Classifier | Agreement metrics | | | | Classifiers | Agreement metrics | | | |
|---|---|---|---|---|---|---|---|---|---|
| | k | PA. | SD | MD | | k | PA | SD | MD |
| (SO)Senti4SD | .83 | 86% | 1% | 12% | Senti4SD vs SentiCR | .77 | 83% | 3% | 14% |
| (SO)SentiStrengthSE | .74 | 80% | 2% | 18% | Senti4SD vs SentiStrengthSE | .79 | 84% | 2% | 15% |
| (SO)SentiCR | .76 | 82% | 3% | 15% | SentiCR vs SentiStrengthSE | .73 | 80% | 3% | 17% |
| (SO)SentiStrength (baseline) | .77 | 82% | 3% | 15% | | | | | |
| (Jira)Senti4SD | .67 | 83% | 0 | 17% | Senti4SD vs. SentiCR | .76 | 88% | 0 | 12% |
| (Jira)SentiStrengthSE | .70 | 83% | 0 | 17% | Senti4SD vs. SentiStrengthSE | .70 | 83% | <1% | 16% |
| (Jira)SentiCR | .73 | 86% | 0 | 14% | SentiCR v. SentiStrengthSE | .81 | 89% | <1% | 10% |
| (Jira)SentiStrength (baseline) | .48 | 66% | <2% | 33% | | | | | |

## 4 RESULTS

**RQ1: To what extent do different SE-specific sentiment analysis tools agree with emotions of software developers?**
For both Stack Overflow and Jira datasets the SE-specific tools outperform the SentiStrength baseline. Higher classification accuracy is observed when trained with gold standard of SE domain. All SE-specific tools are able to correct misclassification of neutral texts as either positive or negative.

**RQ2: To what extent do results from different SE-specific sentiment analysis tools agree with each other?**
Substantial to perfect agreement is observed for all couples of the tools. Strong disagreement is never observed in Jira dataset and is <3% in Stack Overflow.

**RQ3: To what extent do the labeling approach (model-driven vs. ad hoc annotation) has an impact on the performance of SE specific sentiment analysis tools?**
Compared to the model-driven annotation, they observe reduced performance for both datasets in case of ad-hoc datasets. F-measure macro averages indicate lower performance. However, strong disagreement is very low.

## 5 DISCUSSION

SE-specific tuning may improve the accuracy of sentiment classifiers in comparison to off-the-shelf sentiment analysis tools. However, theoretical models of effect should be taken into consideration. These tools show high agreement with respect to manual labeling and also with each other. They have done manual error analysis of the mis-classified texts and report following lessons:

- Reliable sentiment analysis in SE is possible.
- Tuning of tools in the SE domain enhances accuracy.
- Preliminary sanity check is always recommended.
- Grounding research on theoretical models of affect is recommended.

## 6 CONCLUSION

This is a study of benchmarking the performance of sentiment analysis tools for software engineering domain. If manual annotation of gold standard is based on theoretical models of affect, then sentiment analysis is possible in SE. Custom retraining of classifiers with appropriate SE specific gold standard. Unsupervised approach based on lexicon suffices if supervised training is not possible. They plan to extend the contributions of this study by performing a cross-study between datasets as a further performance evaluation technique.

## REFERENCES

[1] P. Jongeling, S. Datta Sarkar, and A. Serebrenik. 2017. On negative results when using sentiment analysis tools for software engineering research. In *Empirical Software Engineering,*. ESE, 2543–2584.
[2] N. Novielli, D. Girardi, and F. Lanubile. 2018. A Benchmark Study on Sentiment Analysis for Software Engineering Research. In *Proceedings of 15th International Conference on Mining Software Repositories*. ACM, 12.